Leong & Mihalcea: Measuring the Semantic Relatedness Between Words and Images

Seminar: Distributionelle Semantik jenseits der Wortbedeutung (Matthias Hartung)

Michael Haas, haas@cl.uni-heidelberg.de

22-07-2013

Overview

- Introduction Multimodal Semantics
- Algorithm: Text + Pictures
- Results
- **Questions? Too fast? Ask!**

Multimodal Semantics

- Distributional Semantics on text corpora: uni-modal
- Integrate different modalities: multi-modal
  - Feature Norms
  - Pictures
- Why:
  - Obvious things go un-mentioned
  - Human cognition is situated

  → Distributional semantics is like "learning meaning by listening to the radio"[1]

[1] McClelland, cited according to Johns &Jones, 2011

Algorithm: Text + Pictures

- Task: measure semantic relatedness between words and images
- Data Set: ImageNet, extension of WordNet
  - Select 167 synsets
  - Select nouns from synsets and glosses
  - Select one image at random from synset
- How to compare images and words?

Algorithm: Representation

- For text: build term-document matrix
  - Vector length: 167 documents
- For images: represent image as bag of visual words

Algorithm: Bag of visual words

- General approach for feature extraction from images
  - Feature Detection: split image into partitions
  - Feature Description: represent image as set of vectors
  - Visual Codeword Generation: cluster vectors

Algorithm: Bag of visual words

- Extract 20px square patches at every 10px boundary
- Represent using SIFT descriptors: Scale-Invariant Feature Transform
- Cluster into 1000 code words
  $\rightarrow$ Image is now represented as a bag of visual code words

CMSM for Sentiment Analysis: Eval Results

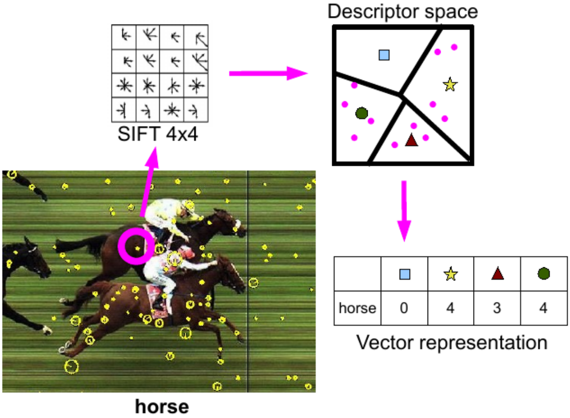

Figure : Bruni et al., 2012

Algorithm: Map images into document space

- Represent each code word as vector: distribution over document space
  $\rightarrow$ Image is represented as set of vectors
- Flatten image represention: sum over all vectors
  $\rightarrow$ Image is now represented as a single vector in document space

Algorithm: Compare images and words

- Words and images are mapped into document space
- Reduce dimensions using LSA
- Measure similarity: cosine similarity
  $\rightarrow$ Direct comparison of vectors in *term-document* and *codeword-document* space

Evaluation

- Image-Centered Scenario
  $\rightarrow$ Given 12 associated words, rank according to relatedness to image
- Arbitrary-Image Scenario
  $\rightarrow$ Measure similarity between arbitrary images and words irregardless of synset membership
- Gold Standard: extract 12 words from synset, relatedness rated by MTurkers

Evaluation: Baselines

- Random baseline
- Vector-based baseline w/o LSA
- Upper bound: human performance based on annotator data

Evaluation: Results

- Image-Centered
  - Vector-based baseline: 0.262 correlation to gold standard
  - LSA-based: 0.339
  - Human upper bound: 0.687
- Arbitrary-Image
  - Vector-based: 0.291
  - LSA-Based: 0.353
  - Human upper bound: 0.764
- Adding more synsets brings correlation values to ~ 0.45

Summary

- Comparing images to text: it works!
- More data is better data
- How can we enrich textual data with image data?
  $\rightarrow$ For starters, just concatenate textual vector and pictoral vector (Bruni et al., 2012)

References I

- Leong, C. W., & Mihalcea, R. (2011, January). Measuring the semantic relatedness between words and images. In Proceedings of the Ninth International Conference on Computational Semantics (pp. 185-194). Association for Computational Linguistics.
- Bruni, E., Boleda, G., Baroni, M., & Tran, N. K. (2012, July). Distributional semantics in technicolor. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 (pp. 136-145). Association for Computational Linguistics.